BBN CTS English System

Spyros Matsoukas, Rukmini Iyer, Owen Kimball,
Jeff Ma, Thomas Colthurst, Rohit Prasad, Chia-Lin Kao

Outline

- Decoding strategy
- Speaker Adaptive Training (SAT)
- Keeping low-count n-grams
- New training transcripts and data
- Part of Speech (POS) Language Model
- Lattice MLLR
- MMI experiments
- System Combination

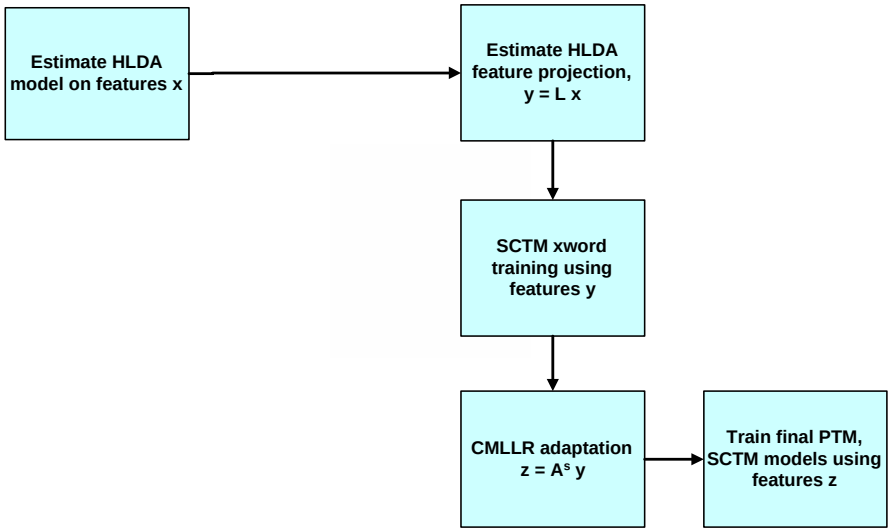
Decoding Strategy

- Feature extraction
  - VTLN, PLP, cepstral mean and covariance normalization
- Pass 1
  - ML GI PTM non-crossword triphone models in forward pass, bigram LM
  - ML GI SCTM non-crossword quinphone models in backward pass, approximate trigram LM
  - ML GI SCTM crossword quinphone models in N-best rescoring pass, full trigram LM
- Pass 2
  - HLDA + CMLLR adaptation
  - MLLR adaptation with 2 regression classes
  - MMI GI HLDA-SAT models in backward and N-best rescoring passes
- Pass 3
  - Lattice MLLR adaptation
  - POS-smoothed LM N-best rescoring

Speaker Adaptive Training

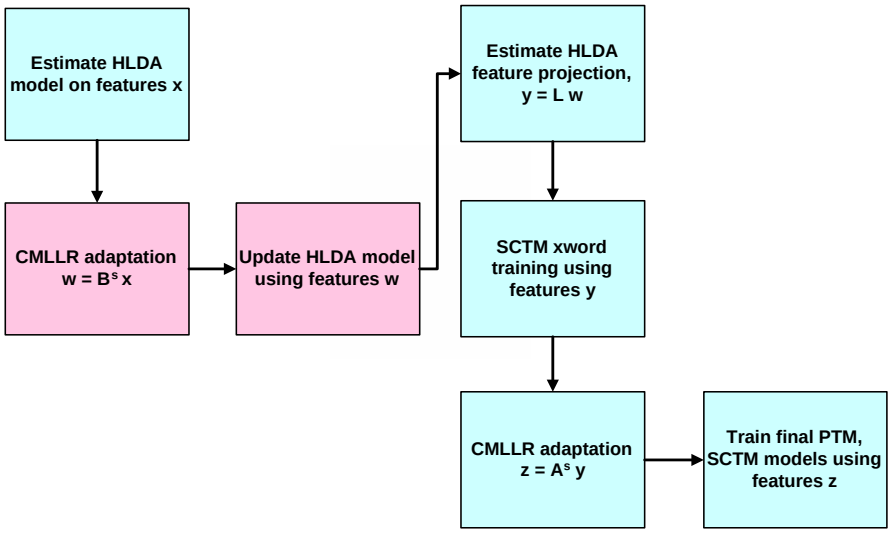
- Explored the use of Constrained MLLR (CMLLR) for SAT
  - Our regular SAT performs model mean transformations through MLLR. This complicates mean update equations for both ML and MMI.
  - CMLLR SAT performs feature transformations in reduced feature space (after HLDA is applied), making it easy to integrate with regular ML and MMI training
- Developed HLDA SAT
  - HLDA SAT extends CMLLR SAT by estimating additional speaker dependent feature transforms in the original feature space (before HLDA is applied)
  - This reduces the overlap between the HLDA classes, resulting in better subspace projections

CMLLR SAT



BBN TECHNOLOGIES
A Verizon Company

HLDA SAT



BBN TECHNOLOGIES
A Verizon Company

SAT Results

Results on Eval01, using ML GI models, after two adaptation passes

CMLLR-SAT	WER
no	25.8
yes	25.3

- **Tried both CMLLR SAT and HLDA SAT on three domains:**
  - **CTS English, BN English, CTS Mandarin**
  - **Modest gain (0.5%) from CMLLR SAT (compared to SI) on both CTS and BN English**
  - **Additional gain (0.7%) from HLDA SAT on BN English and CTS Mandarin, but no gain on CTS English**

Keeping Low-count N-grams

- **In January, we discussed the effect of keeping all n-grams, including those with a single training example**
- **We changed last year's 3-gram LM (trained on Swbd plus BN), to keep all 3-grams**
- **Result for Dev03**
  - **Dev03 is the Swbd 2 Phase 2 + Swbd-Cellular parts of Eval01**

System	WER
Use n-grams cutoffs	28.3
Keep all n-grams	27.8

Transcriptions in RT-03 CTS System

- Switched from LDC/MSU to CU transcripts
- CU uses more consistent text conventions and corrected segmentation
- No non-speech words in CU transcripts
  - We increased Gaussians/codebook for SILENCE phoneme to model non-speech as well
- No significant difference in performance
  - We used CU transcripts since they were cleaner

Transcribing More Training Data

- Previous gains for using 20 hours of cell data encouraged us to seek more in-domain training
- Investigated fast transcription using local service, CTRAN
  - BBN derived utterance-level segmentation
  - Segmentation process also rejects segments with many errors or poor segment boundaries
- In January, we showed training on 20 hours of Swbd1 using original LDC, CTRAN, or fast LDC transcripts gave comparable results
  - CTRAN was 0.5% worse than original LDC
  - Having auto segmenter make shorter segments reduces difference to 0.1%

Adding New Training Data

- We then had CTRAN transcribe 100 hours of Switchboard 2 data
  - 50 hours Cellular; 50 hours Switchboard 2 Phase 2
  - 80 hours survived post-processing rejection
- Added to existing 295-hr training set
- WER on Dev03 with ML models

LM Transcripts	AM Transcripts	WER
CU+BN	CU	27.8
CU+CTRAN+BN	CU	27.5
CU+CTRAN+BN	CU+CTRAN	25.8

More Language Modeling Data

- Using Additional Text Resources
  - 141M words from Broadcast News (BN), epoch 1992-1996
  - 64M words of web text (WB), cleaned and normalized by University of Washington
  - 47M words of archived text (AR) from CNN and PBS broadcasts, epoch 2000-2002
- Train similarity-weighted language models
  - Weight BN articles based on similarity to Swbd 1 corpus
  - Empirical weight for all of the WB data
  - Weight AR articles based on similarity to Fisher topic descriptions

Using Additional LM Data

- Results on Dev03, 35K vocabulary, after one adaptation pass, ML models

LM Transcriptions	WER
CU+CTRAN+BN	25.8
CU+CTRAN+BN+WB	25.4
CU+CTRAN+BN+WB+AR	25.3

- Augment decoding vocabulary -- Results on Eval01 after two adaptation passes, MMI models

Vocabulary Size	WER
35K	21.4
45K	21.3
55K	21.1

Part-of-Speech Language Model

- Continue using part-of-speech (POS) LM to rescore N-best

$$P(w \mid \text{history}) = \sum_{i=1}^{N} P(c_i \mid \text{history}) P(w \mid c_i, \text{history})$$

- Automatically tag all new text resources using Adwait Ratnaparkhi's part-of-speech tagger

  ftp.cis.upenn.edu/pub/adwait/imx/imx.tar.gz

- Results on Eval01 after two adaptation passes

Language Model	WER
Trigram	21.1
POS Trigram	20.8

Lattice MLLR adaptation

- **Used lattice MLLR* adaptation for the crossword SCTM MMI models in the final decoding pass**
  - **lattices were generated from N-best**
  - **used a maximum of 128 regression classes (5 on average)**

Adaptation Method	WER
1-best MLLR	20.8
Lattice MLLR	20.5

(*) L. F. Uebel and P. C. Woodland, "Improvements in Linear Transform Based Speaker Adaptation," ICASSP 2001

MMI Overview

- **Since RT-02, added lattice-based MMI**
- **At Jan 2003 workshop, reported MMI experiments with model size, smoothing, compound words, lattice quality**
- **Showed 4.3% relative improvement over best ML**
- **Relative gain has increased to 7.2% over best ML in final system (20.7% vs. 22.3% on Eval01)**
  - **Mainly due to the use of more training data**
- **Tried to improve MMI further, but without success**
  - **frame weighting**
  - **I-smoothing**

MMI: Frame weighting and I-smoothing

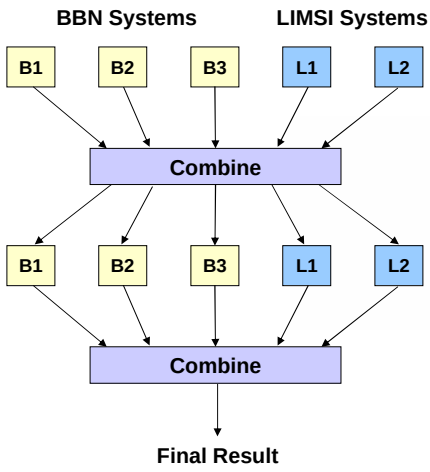
- **Frame weighting**
  - **In discriminative training, data near the decision boundary is more important. Therefore, try weighting it more, using an exponentiated raised cosine function of the posterior probability of the reference states in the denominator lattice.**
  - **0.5% gain on Eval01, using non-crossword models**
  - **0.2% gain using crossword**
  - **gain disappeared after integration with other improvements**
- **I-smoothing***
  - **tried several values of τ, observed no gain over regular MMI**

(*) D. Povey and P. C. Woodland, "Minimum Phone Error and I-smoothing for Improved Discriminative Training", ICASSP 2002

Systems for Transatlantic Combination

- **BBN systems for combination**
  - **B1: PLP, MMI, GI, 49 phonemes**
  - **B2: PLP, MMI, GI, 41 phonemes**
  - **B3: MFCC, MMI, GI, 41 phonemes**
- **LIMSI systems**
  - **L1: PLP-S: short-term cepstral mean and variance normalization**
  - **L2: PLP-R: reduced phone set models (35 instead of 45)**

CTS English Transatlantic System Architecture



Adapt to combined result and redecode

BBN TECHNOLOGIES
A Verizon Company

Transatlantic (TA) Combination results

System	Before TA Adapt		After TA Adapt	
	Eval01	Eval02	Eval01	Eval02
B1	20.5	24.4	19.0	22.9
B2	20.7	24.2	19.0	22.7
B3	21.2	24.8	19.3	22.9
L1	21.9	25.6	20.3	24.0
L2	21.8	25.6	20.5	23.9
Combo	18.4	21.9	18.0	21.3

Eval01 results use manual segmentation, Eval02 use automatic

BBN TECHNOLOGIES
A Verizon Company

CTS Progress since RT-02

- Progress this year on Eval02 and Ears Progress set

System	Eval02 WER
BBN RT-02, manual segmentation	28.4%
BBN+LIMSI RT-03, auto segmentation	21.3%
Relative Reduction	25.0%

System	Progress Test WER
BBN RT-02, MIT-LL auto segmentation	27.8%
BBN+LIMSI RT-03, auto segmentation	17.5%
Relative Reduction	37%

- RT-03 Current Test Results, BBN+LIMSI System

Fisher	Swbd 2	Overall
16.7	23.8	20.4

Conclusions

- Significant gains this year for
  - MMI
  - New in-domain acoustic data (CTRAN)
  - Additional LM data
  - Keeping all n-gram counts
  - Lattice MLLR
  - New HLDA-SAT: No help for CTS English, but see Mandarin CTS, English BN results
  - System combination
- Transatlantic ROVER + Adaptation gives effective integration with LIMSI system

BBN CTS Mandarin System

Jeff Ma, Spyros Matsoukas, Thomas Colthurst, Owen Kimball, Rukmini Iyer, Chia-Lin Kao, Dongxin Xu, Irfan Karadag

Outline

- Systems overview
- Improvements since 2001 Eval system
- Mixture exponents in BBN system
- Creating a CallFriend Dev Set
- System combination

Systems Overview

- **2001 Evaluation System used**
  - MFCC analysis
  - mixture exponents in training and decoding
  - three MLLR adaptation passes alternating between SAT and SI models
  - finishing with adapted decodes on features analyzed at the three frame rates of 100, 125, and 80 frames per second
  - approximately 157,000 gaussians in best models
- **"Modern" systems used**
  - PLP analysis
  - fuzzy labels in training
  - no mixture exponents in training or decoding
  - one pass of HLDA+CMLLR+MLLR adaptation
  - one pass of lattice MLLR adaptation
  - approximately 47,000 gaussians in best models

Mandarin CTS Improvements

System	Unadapted CER	Best Adapted CER
2001 Evaluation System	51.8	47.0
Bug Fixes, Better PTM Models	50.5	46.7
PLP Analysis	49.6	46.0
Modern Training, CMLLR-SAT	50.4	46.3
HLDA-SAT	50.4	45.7
BN in LM	50.3	45.2
Mixture exponents and 143K Gaussians	49.2	43.7
MMI with exponents	-	42.0

All CERs measured on the 1997 Callhome Mandarin evaluation set.

Mixture exponents

- **Using mixture exponents helps smooth large models trained on small amounts of data**
  - **observation probability of x without mixture exponents**

$$p(x) = \sum_k w_k g_k(x) = g_b(x) \sum_k w_k \frac{g_k(x)}{g_b(x)}$$

$$g_b(x): \text{top-1 Gaussian density value}$$

  - **observation probability of x with mixture exponents**

$$\hat{p}(x) = [g_b(x)]^{\alpha} \sum_k w_k^{\beta} \left(\frac{g_k(x)}{g_b(x)} \right)^{\gamma}$$

Creating a CTS Mandarin Dev Set

- **We wanted a CallFriend (CF) dev set to match Eval data**
- **SRI defined a dev set drawn from existing transcribed CF data**
  - **These calls are ½ hr each and are part of our training.**
  - **Using dev set required either dropping those conversations from training (1/2 the CF training) or dropping just the 5-minute test selections and keeping test speakers in training.**
- **Instead created new 1 hour set from untranscribed CF data**
- **Untranscribed data:**
  - **19 conversations; 18 from mainland, one from Taiwan (excluded)**
  - **Only 11 male speakers in the 18 conversations**
- **Dev set used 12 conversations, including 13 females and all 11 males**
  - **Selected 5 minutes from each conversation**

Creating a CTS Mandarin Dev Set (cont'd)

- **Transcribed using BBN tool**
  - **Followed LDC rules, though not as careful as LDC (speed was important)**
- **Dropped one conversation side due to transcription errors, resulting in a 55 minute, 23 speaker test set**
- **Distributed to community 1 week before eval (but < 1 day after transcription)**

Mandarin CTS System Combination

System(s)	Front-End	Mixture Exponent	Phoneme Set Size	CER on CF Dev Set	CER on Eval97 Set
M1	PLP	Yes	85	42.8	43.7
M2	PLP	No	85	43.4	43.7
M3	PLP	Yes	148	43.4	44.1
M4	MFCC	Yes	85	43.4	44.4
M1+M3+M4				40.7	-
M2+M3+M4				40.0	-
M1+M2+M3+M4				39.9	41.5

All models are MMI.
All combinations were done with simple ROVER.
RT03 current test set result: 42.7%

BBN CTS Arabic System

Thomas Colthurst, Spyros Matsoukas, Jeff Ma, Owen Kimball, Rukmini Iyer, Chia-Lin Kao, Will Seitz, John Makhoul

Egyptian Arabic CTS Improvements

System	Best Adapted WER
S1. Baseline (July 2002)	51.4
S2. New binaries (Feb 2003)	51.3
S3. Slow VTL	51.2
S4. PLP Analysis	50.9
S5. Modern ML training, no exponents	51.9
S6. Modern ML training	50.8
S7. Modern MMI training, no exponents	51.6
S8. Modern MMI training	50.5

All WERs measured on the 1997 Callhome Arabic evaluation set.

Arabic CTS System Combination

Systems	WER Range	Combined WER
S1 @ 100, 125, 80 f/s (2002 baseline)	51.4 - 54.1	49.9
S3, S4	50.9 - 51.2	48.9
S3, S4, S5, S7, S8	50.5 - 51.9	47.6
S3, S4, S7, S8	50.5 - 51.6	47.6 *
S4, S7, S8	50.5 - 51.6	47.9

All Combinations were done with simple ROVER.

** = 2003 Evaluation system configuration, obtained 37.5% on RT03 evaluation test set.*

Mysteries of Arabic CTS

- **Only 0.3% absolute gain from MMI**
- **No gain from using crossword models**
- **No gain from adding the 20 extra conversations in the LDC 2002 supplemental release to our training**